Detecting Overlapping Link Communities by Finding Local Minima of a Cost Function with a Memetic Algorithm Part 1: Problem and Method

Frank Havemann* Jochen Gläser† Michael Heinz‡

Abstract

We propose an algorithm for detecting communities of links in networks which uses local information, is based on a new evaluation function, and allows for pervasive overlaps of communities. The complexity of the clustering task requires the application of a memetic algorithm that combines probabilistic evolutionary strategies with deterministic local searches. In Part 2 we will present results of experiments with citation networks.

1 Introduction

Communities in networks are commonly defined as cohesive subgraphs which are well separated from the rest of the network. This vague concept of communities is operationalised in a variety of ways (Fortunato 2010). The utility of algorithms for the detection of communities in networks partly depends on their 'conceptual fit', i.e. on the degree to which they match properties of the phenomenon that is represented (Hric, Darst, and Fortunato 2014). Achieving such a conceptual fit may require unusual combinations of ideas from network analysis, as is the case with the question and the algorithm presented in this paper.

Consider the following three properties of a network and the task of community detection. First, links between nodes contain better information about communities than the nodes that are to be clustered. In this case, link clustering appears to be the method of choice. Constructing communities by clustering links has been proposed by Evans and Lambiotte (2009) and by Ahn, Bagrow, and Lehmann (2010) as a method for the construction of overlapping communities of nodes. In addition, clustering links is likely to be advantageous whenever the information asymmetry described above occurs, i.e. whenever links rather than nodes have the real-world properties whose similarity shall be reflected by clusters.

Second, overlapping communities must be a possible outcome of the algorithm because the real-world phenomenon under investigation is known to have such a structure. For the same reason, pervasive overlaps must be possible, i.e. overlaps that extend to all nodes rather than just the boundary nodes of a community. The construction of overlapping communities is by now a well-known and frequently addressed problem of network analysis (Fortunato 2010; Xie, Kelley, and Szymanski 2013; Amelio and Pizzuti 2014).

Third, the phenomena to be represented by communities are local in that they emerge from local interactions represented by neighbouring nodes and links in the network. If this is the case, the use of local rather than global information may return better communities and a better community structure of the network (Clauset 2005; Lancichinetti, Fortunato, and Kertesz 2009; Havemann, Heinz, Struck, and Gläser 2011).

All three ideas have been developed in network analysis. However, as reviews of algorithms indicate (Fortunato 2010; Xie et al. 2013; Amelio and Pizzuti 2014), link clustering, pervasively overlapping communities and use of local information have not yet been combined all three, possibly because the task for which this is necessary has not yet arisen.[1]

*Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin, D 10099 Berlin, Dorotheenstr. 26 (Germany)

†Center for Technology and Society, TU Berlin (Germany)

‡Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin

[1]The only apparent exception is the work by Lei Pan *et al.* which, however, compromises in two expects, global

There is at least one task for which this combination of link-based approach, pervasive overlaps and local approach is necessary, namely the detection of thematic structures (topics) in networks of papers.

In networks of papers and their cited sources, citation links (links between a publication and the sources it cites) are thematically more homogenous than nodes (papers), and thus provide better information for clustering, than the papers themselves. While papers commonly belong to more than one scientific topic, many citation links can be assumed to be homogenous in that the link between paper and source belongs to only one topic. If it belongs to more than one topic these topics often are not very distant from each other.

Scientific topics are known to overlap pervasively, which means that their reconstruction as communities of papers must reflect this pervasive overlap. Topics are also locally emergent phenomena in that they represent coinciding and mutually referring perspectives of researchers (the authors of the papers).

In order to reconstruct scientific topics from networks of papers and their cited sources, then, we need an algorithm that clusters links, can construct pervasively overlapping communities, and uses mainly local information. In this paper, we present such an algorithm (in Part 1) and its application to citation networks (in Part 2). We propose a local cost function for the independent evaluation of each link community by relating its external to its total connectivity in the network. The cost function is almost completely based on local information, the only global information used is the number of links in the whole network. The independent evaluation of each subgraph with a local cost function means that communities can be constructed independently from each other, which enables pervasive overlaps.

The cost function we propose for subgraph evaluation is solely based on the network's topology and not on link similarity. Generally, clustering by optimising a (global or local) evaluation function needs no measure of similarity of clustered elements but results in clusters the elements of which are seen as similar in some sense. In contrast, the approach to link clustering proposed by Ahn et al. (2010) is based on link similarity. The authors estimate the similarity of two links by comparing their sets of neighbouring nodes. This is not very appropiate for citation links because we would estimate thematic similarity of thematically nearly homogenous elements (citation links) with sets of very inhomogenous elements (papers, cited sources). In the case of citation networks, it would be better to measure link similarity by using textual information from citing and cited documents.

The local construction of topics, their varying size and pervasive overlaps make it likely that topics form a poly-hierarchy i.e. a hierarchy where a smaller topic can be a subtopic of two or more larger topics that have no hierarchical subtopic relation. This poly-hierarchy of topics should be reflected in a poly-hierarchy of communities.

Communities without sub-communities can be well separated and very cohesive, too, but inside larger communities there can exist well separated sub-communities which diminish the cohesion of their super-community.

Since the cost landscape of link communities has many local minima, purely deterministic search strategies are not efficient. This is why we designed a memetic search that combines an evolutionary algorithm with deterministic adjustments in the cost landscape. Evolutionary algorithms have already been used for identifying communities in networks (Fortunato 2010, p. 106). Some authors have even applied evolutionary algorithms to link clustering but all used global evaluation functions (Pizzuti 2009; Li, Zhang, Wang, Liu, and Zhang 2013; Shi, Cai, Fu, Dong, and Wu 2013). Memetic evolutionary algorithms have also been applied to reconstruct communities but only for node clustering and only with global evaluation functions (Gong, Fu, Jiao, and Du 2011; Pizzuti 2012; Gach and Hao 2012; Ma, Gong, Liu, Cai, and Jiao 2014).

2 Strategy

The strategy we apply in response to the three challenges described in the introduction consists of three main steps. We develop an evaluation function for link communities that uses local information. This evaluation function makes it possible to construct each community indepen-

information used in the end and no pervasive overlap because link clusters are disjunct (Pan, Wang, Xie, and Liu 2011; Pan, Wang, and Xie 2012). Furthermore, they differ from our approach because they propose an evaluation function for link clustering which is derived within the node clustering approach.

dently from all others, which in turn enables pervasive overlaps because inner links (links all of whose neighbours are community members) of one community can also be inner links of another community. We then design an algorithm that constructs local communities.

For the first step, we followed a suggestion by Evans and Lambiotte (2009) to obtain link clusters by clustering vertices in a network's line graph. We defined a local cost function $\Psi(L)$ in the line-graph approach which we call *ratio node-cut*. It can be used to identify link communities by finding local minima in the cost landscape. Since $\Psi(L)$ evaluates the boundary between a subgraph and the rest of the network, communities can be constructed independently of all other communities.

The cost landscape of $\Psi(L)$ is often very rough i.e. has many local minima that may correspond to very similar subgraphs. Therefore, the resolution of the algorithm must be defined by setting a minimum distance (number of links that differ) between subgraphs corresponding to different local minima. We define the range of a community as a distance in which no subgraph exists that has a lower Ψ-value.

Since the task of finding communities in large networks is always very complex, heuristics must be applied. This applies even more strongly to link clustering because networks contain many more links than nodes, and particularly to the rough Ψ-landscape. We chose an evolutionary algorithm but accelerate evolution by combining it with a deterministic local search in the cost landscape. This approach is called memetic (Neri, Cotta, and Moscato 2012). Memetic algorithms can also find local optima of a local cost function (Vitela and Castaños 2012).

In evolutionary algorithms, individuals occupy places in the cost (or fitness) landscape. In our local algorithm, populations are sets of different subgraphs. We start with a random initialisation of the population of some definite size. The genetic operators of crossover, mutation, and selection are repeatedly applied to move the population into optima. In memetic algorithms each crossover and each mutation is followed by a local search.

In large networks exploring the cost landscape by adding or removing individual links is very time-consuming. We therefore begin the search with a coarse search phase that adds or removes groups of links by adding or removing nodes with all their links, and follow it with fine search phase, namely link-wise memetic evolution or at least a link-wise local search.

3 The cost function: *ratio node-cut*

3.1 Node-induced and link-induced subgraphs

Traditionally, the boundary of a community is drawn between nodes and therefore cuts the links between nodes inside and outside the community. If we consider communities as clusters of links rather than nodes, the perspective must be reversed. While the boundary of a node community cuts links, the boundary of a link community cuts nodes.

A node community is a connected subgraph defined by a node set C. It contains all links existing between nodes in C. A link community is a connected link-induced subgraph. It contains all nodes attached to links of a given set L. There can be links existing between a link community's nodes which are not in L.

Cost functions of a subgraph can be defined by relating a measure of external to a measure of total connectivity. This ratio should be minimal for well separated and cohesive subgraphs i.e. for communities.

Node communities can be defined as connected subgraphs corresponding to minima in cost landscapes where places correspond to node-induced subgraphs. Correspondingly, link communities can be defined as connected subgraphs corresponding to minima in cost landscapes where places correspond to link-induced subgraphs.

In the following, we only consider connected unweighted graphs $G = (V, E)$. The number of edges (or links) is $m = |E|$, the number of vertices (or nodes) is $n = |V|$. With k_i we denote the degree of node i. The internal degree of node i, denoted by $k_i^{\text{in}}(L)$, is the number of links attached to node i which are in link set L. The external degree of node i is $k_i^{\text{out}}(L) = k_i - k_i^{\text{in}}(L)$.

3.2 External connectivity

We first consider measures of *external connectivity* of a subgraph which are useful for constructing node or link communities. The simplest measure of external connectivity of a node-

induced subgraph is the *cut size* that equals the sum of weights of boundary links i.e. the links connecting the subgraph with the rest of the graph (Fortunato 2010, p. 92). If link weights represent electrical conductance, cut size measures the total conductance of all boundary links. Cut size can be calculated as the sum of external degrees $k_i^{\text{out}}(L)$ of boundary nodes (subgraph members with boundary links).

Applying these considerations to the external connectivity of a link-induced subgraph leads to a simple measure of external connectivity as the sum of $k_i^{\text{out}} k_i^{\text{in}}/k_i$ of boundary nodes:

$$\sigma(L) = \sum_{i=1}^{n} \frac{k_i^{\text{out}}(L) k_i^{\text{in}}(L)}{k_i}. \tag{1}$$

Only for boundary nodes of L we have $k_i^{\text{out}} k_i^{\text{in}} > 0$. That means, we can restrict the sum in the formula to boundary nodes. In function $\sigma(L)$ the external degrees k_i^{out} are weighted with subgraph membership-grade k_i^{in}/k_i of the boundary nodes. The function $\sigma(L)$ can be derived from the total conductance or cut size of link sets in the graph's line graph if the line graph's edges are weighted with $1/k_i$—a weighting proposed by Evans and Lambiotte (2009). The derivation can be found in Appendix A.

Each term of $\sigma(L)$ equals the conductance of a boundary node i i.e. the total conductance for currents flowing out of the subgraph through this node. We call $\sigma(L)$ the *node cut* of a link-induced subgraph.

3.3 Internal and total connectivity

Now we discuss measures of *internal and total connectivity* of subgraphs induced by node and by link sets, respectively. In the case of node-induced subgraphs $k_{\text{in}}(C) = \sum_{i \in C} k_i^{\text{in}}(C)$ is an appropriate measure of internal connectivity of node set C. Total connectivity of C is then the sum of degrees of all nodes in C:

$$k_{\text{total}}(C) = \sum_{i \in C} k_i^{\text{in}}(C) + k_i^{\text{out}}(C) = \sum_{i \in C} k_i. \tag{2}$$

For a link-induced subgraph we can use the sum of internal degrees, weighted with their membership, as a measure of internal connectivity:

$$\tau(L) = \sum_{i=1}^{n} \frac{k_i^{\text{in}}(L) k_i^{\text{in}}(L)}{k_i}. \tag{3}$$

The sum is restricted to nodes attached to links in L because other nodes have $k_i^{\text{in}}(L) = 0$. Total connectivity of L is then given by the sum $\sigma(L) + \tau(L) = \sum_{i=1}^{n} k_i^{\text{in}}(L) = k_{\text{in}}(L) = 2|L|$. The derivation can be found in Appendix A.

3.4 Cost function

Relating external to total connectivity leads us to cost functions whose minima correspond to well separated and cohesive subgraphs. On the other hand, we also achieve a size normalisation when we divide external by total connectivity. This is welcome, because the boundary length (measured by external connectivity) tends to increase with size (here measured by total connectivity $k_{\text{in}}(L) = 2|L|$)—at least for not too large subgraphs in not too small networks. If a subgraph occupies more than one half of the network its boundary tends to become shorter with increasing size. A simple size normalisation that accounts for the finite size of the network is achieved by adding to the external-total ratio of a subgraph the same ratio of its complement. For small subgraphs in a large network the second ratio is very small. For node-induced subgraphs this normalisation was introduced by Wei and Cheng (1989) and named *ratio cut*. For link-induced subgraphs we analogously define a cost function *ratio node-cut* as

$$\Psi(L) = \frac{\sigma(L)}{k_{\text{in}}(L)} + \frac{\sigma(E\backslash L)}{k_{\text{in}}(E\backslash L)} \tag{4}$$

$$= \frac{\sigma(L)}{k_{\text{in}}(L)(1 - k_{\text{in}}(L)/2m)}. \tag{5}$$

The expression on the r.h.s. is obtained because $\sigma(E\backslash L) = \sigma(L)$ and $k_{\text{in}}(E\backslash L) = 2m - k_{\text{in}}(L)$. Ratio node-cut Ψ is not strictly local but the only global information needed here is the total number of links m. In the limit of small subgraphs in large networks we achieve approximately strict locality because we have $k_{\text{in}}(L) \ll 2m$ and we therefore obtain

$$\Psi(L) \approx \frac{\sigma(L)}{k_{\text{in}}(L)}, \tag{6}$$

which equals a strictly local cost function for the construction of link communities introduced by us earlier (Havemann et al. 2012). Our cost function $\Psi(L)$ rewards separation of link community L but not really its cohesion. Yang and Leskovec (2012) found that evaluating node subgraphs with *conductance*—a measure analogue to $\sigma(L)/k_{\text{in}}(L)$ in the world of

node communities—can also lead to communities with low cohesion.

That means, using cost function Ψ we emphasize separation and require only a minimal cohesion of subgraphs, which is expressed by the demand that subgraphs must be connected. Otherwise an unconnected subgraph with parts in very different regions of the network could be a community.

Function $\sigma(L)$ vanishes for the empty subgraph with $L = \emptyset$ and for the full graph with $L = E$. In both cases, the denominator of the cost function also vanishes and we obtain zero divided by zero but it makes sense to define $\Psi(E) = \Psi(\emptyset) = 1$ because Ψ of one link (1, 2) with vanishing weight w_{12} approximates 1:

$$\Psi((1,2)) = w_{12}\frac{(k_1 - w_{12})/k_1 + (k_2 - w_{12})/k_2}{2w_{12}(1 - 2w_{12}/2m)}. \tag{7}$$

Our cost function is symmetric: $\Psi(L) = \Psi(E\backslash L)$, i.e. the cost function is the same for a link-induced subgraph and the subgraph induced by the complementary link set $E\backslash L$.

3.5 The cost landscape

Each place in the cost landscape represents a link-induced subgraph. Two places in the landscape have a direct relation if and only if the corresponding subgraphs differ in one link. The height of each place is given by the value of the cost function $\Psi(L)$. The global minimum of the cost function is reached for a division of the set E of all links that produces the two best link communities in terms of separation. As a simple example, we determined the Ψ-landscape of the bow-tie graph (Figure 1, for calculations see Appendix B). We expect a cut through the central node to be the best division in two link communities (the two triangles). Indeed, the landscape has two minima with $\Psi = 1/3$, which correspond to the two triangles. There are no further local minima.

We do not restrict the search for link communities to finding only the global minimum but define a link community as a connected link-induced subgraph which corresponds to any local minimum in the Ψ-landscape. Since the Ψ-landscape of larger graphs contains many local minima, we need a filter to select the locally best link communities. For this reason, we restrict our search to those minima with a sufficiently large distance to any lower place in the cost landscape. Thus, we have to define the resolution of the search by defining this minimal distance in the landscape. The appropriate resolution depends on the research question about the phenomenon represented by the network. The extent to which two communities should differ in content (of links) to consider them as different depends on the question asked about communities.

Another place in the cost landscape is reached by adding links to and by removing links from the subgraph corresponding to the starting place. The distance between two places in the cost landscape equals the sum of the number of links we have to add and to exclude. In other words: the distance is the size of the symmetric difference between the two link sets. We define the range of a community as the minimal distance to a subgraph with lower cost. Within a community's range there is no better subgraph. The resolution of a search for communities can be defined as the minimal range of communities that are accepted as valid solutions. Depending on the networks real background, a relative resolution can be more appropriate. That means, we demand that any valid community should have a range which is larger than a certain percentage of its size.

In order to determine the range of a community we would need to know its whole environment up to the distance to the nearest lower place in the cost landscape. Otherwise, a lower place only determines an upper bound of the community's range. However, searching the whole environment of a subgraph is practically impossible for large networks. A selective search is necessary, which is why we apply evolutionary and deterministic greedy algorithms. If these algorithms find an upper bound smaller than the set resolution, we can deselect the community. If they don't, the community is provisionally kept

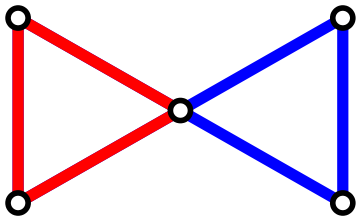

Figure 1: Bow-tie graph

but can later be replaced by a better community within its minimal range defined by the set resolution. We assume, however, that later found better solutions differ only in some links.

4 Memetic search

Memetic algorithms combine random evolution with deterministic local search. In this section, we describe

1. the local search we apply, called adaptation for short,
2. our implementation of the evolutionary approach,
3. the genetic operators of mutation, crossover, and selection we employ in the evolutionary approach.

The memetic algorithm is applied in the search for link communities, which can be done by exploring the cost landscape of a network by adding or removing individual links. For large subgraphs this is very time-consuming. We therefore split the search in a coarse phase, in which we add or remove nodes with all their links to other nodes in the subgraph, and a finer link-wise search, which is applied after communities have been identified by a node-wise search. After communities with a minimal range defined by the set resolution are found in a node-wise memetic search, they are subjected to a link-wise memetic search or at least a link-wise local search.

4.1 Local search

The local search in the cost landscape applies a greedy algorithm for finding local cost minima that correspond to communities. The algorithm starts from the place occupied by the current subgraph and moves to subgraphs with lower Ψ-values. The algorithm is greedy because it always chooses the step that brings the biggest decrease or the smallest increase of Ψ. A step includes or excludes a node with all their links to the nodes already in a subgraph in node-wise local search, and includes or excludes an individual link in link-wise local search.

A valid community can be made invalid and replaced by a better one if the better one is within its minimal range which is set by the resolution parameter. Therefore, the local search has not to find subgraphs with lower cost in each step but can go a number of steps by 'tunneling' through 'barriers' in the landscape (areas with higher Ψ) before reaching lower values which invalidate the community at the tunnel's entrance. Tunneling makes the algorithm more efficient. The maximum length of a tunnel through a barrier of higher Ψ-values is determined by the set resolution.

The local search can begin by a series of either inclusions or exclusions of nodes (links). When no further improvement can be achieved, the search switches from inclusion to exclusion or vice versa. Inclusion and exclusion are continued until no further improvement is possible.

If the exclusion of nodes fragments a subgraph, we proceed with the subgraph's main component. In the link-wise local search the greedy algorithm is allowed to go through intermediary states representing unconnected subgraphs. At the end of the link-wise local search we determine all components of the subgraph. If the subgraph is unconnected we repeat the procedure for each component until we obtain only connected subgraphs with minimal cost.

A greedy algorithm is efficient because the cost reduction for all possible cases of including a neighbour must be calculated only at the beginning of the local search. In the subsequent steps, we only calculate or recalculate cost reductions achieved by adding neighbours of the link (or node) included. Analogously, we proceed when excluding boundary nodes or links (Havemann et al. 2012, Appendix). Otherwise it would be more efficient to include or exclude just the first node (link) which reduces cost.

4.2 Evolution

The general implementation of the memetic algorithm is described by Algorithm 1.[2] The genetic operators of crossover, mutation, and selection (described below) are applied to each generation of communities. Subgraphs generated by crossover and mutation are adapted by a local search. If the starting subgraph is not connected we replace it by its main component. Evolution is terminated when no better best community is found for many generations.

[2]The notation is inspired by a pseudocode given by Merz (2012).

4.3 Genetic operators

Mutation: We mutate a community with mutation variance $v < 1$ by changing maximally a proportion v of its links or nodes. In node-wise memetic evolutions we randomly exclude boundary nodes and then include the same number of neighbouring nodes. In link-wise memetics we experiment with two other mutation operators: we only exclude *or* include links and concentrate changes around one randomly chosen boundary node. (Details can be found in Appendix of Part 2.)

Crossover: From two parent subgraphs we construct two new individuals by taking intersection and union of the subgraphs as starting points for adaptive local searches. Of course, it has no effect to cross such parents where one of them is part of the other one. Normally, evolutionary algorithms include some randomness in the crossover, which in our case would mean to enlarge the intersection by some nodes or links from the union. In contrast, our crossing procedure is deterministic because the boundary of the union of two good communities should also be not too bad. The same holds for the intersection. Deterministic crossover should be (and is) done only once with the same parents. The only random element of our crossover is the random selection of parents.

Selection: From the old population and the results of mutations and crossovers we select the communities with lowest Ψ-values, keeping the population size constant. A new best community is only included if it is inside the minimal range of the best community of the original population. Disregarding the best communities outside the minimal range assures that we do not lose communities which can have a range above the minimum given by the resolution limit we apply. Deselected communities can be used as seeds for other memetic searches.

Renewal: Renewal means to mutate the best community with high variance several times, to adapt the mutants, and to apply a usual selection procedure described above.

Algorithm 1 Pseudocode of memetic evolution for one adapted seed

initialise population **P** by **mutating** the adapted seed with high variance several times and **adapting** mutants
while the best community is not too old **do**
  mutate the best community with low variance and **adapt** the mutants
  if an adapted mutant is new and its cost is lower than highest cost **then**
    add it to population **P**
  end if
  cross the best community with some randomly chosen communities and **adapt** the offspring
  if adapted offspring is new and its cost is lower than highest cost **then**
    add it to population **P**
  end if
  select the best communities so that the population size remains constant
  if there is no better best community for some generations and innovation rate is low **then**
    renew the population by mutating the best community with high variance and **adapt** mutants
    select the best communities so that the population size remains constant
  end if
end while

5 Concluding remarks

In the forthcoming Part 2 of our paper we discuss test results.

Acknowledgement

This work is part of a project in which we develop methods for measuring the diversity of research. The project was funded by the German Ministry for Education and Research (BMBF). We would like to thank the members of the project's advisory group[3] and also all developers of **R**.[4] We thank Alexander Struck for assisting our work and for discussions about the new cost function. We discussed the issue of link similarity in citation networks with Rob Koopman.

[3] `http://141.20.126.172/~div`
[4] `http://www.r-project.org`

Appendix

A Connectivity measures

In this section we derive the connectivity measures for link sets $\sigma(L)$ and $k_{\text{in}}(L)$ from analogue measures in the line graph. We closely follow the arguments given in our earlier paper (Havemann, Gläser, Heinz, and Struck 2012).

We here use $i, j = 1, \ldots, n$ to denote nodes and $k, l = 1, \ldots, m$ for links. With $C(L)$ we name the set of nodes attached to links in the subgraph induced by link set L. If a link k belongs to L its membership $\mu_k(L) = 1$ and zero otherwise.

To construct a network's line graph we first define an auxiliary bipartite graph obtained by putting a node on each link of the original network. The affiliation matrix B of the bipartite graph—also called its incidence matrix—has a row for each of the n original nodes and a column for each of the m original links. Each link column contains only two non-zero elements, namely the elements in the rows of the nodes i and j connected by the link. We can project the bipartite graph back onto the original network with the product BB^{T} which equals its adjacency matrix A (except for the main diagonal).

We obtain the network's line graph by the opposite projection $B^{\mathrm{T}}B$ of the bipartite graph. Evans and Lambiotte (2009) underline, that in all cases of practical intererest the line graph contains the same amount of information as the original network. Knowing $B^{\mathrm{T}}B$ we can almost ever calculate BB^{T} and thus also the network's adjacency matrix A.

Because each node of the original network is represented as a clique in the line graph Evans and Lambiotte (2009) weighted the edges of the line graph with the inverse degree $1/k_i$ of the node i in the original network. They define the line graph's adjacency matrix as

$$E_{kl} = \sum_{i=1}^{n} \frac{B_{ik}B_{il}}{k_i}. \tag{8}$$

Weighting the line graph's edges with the inverse degrees of nodes in the original network is equivalent to an Euclidean normalisation of the nodes' vectors in the affiliation matrix B of the auxiliary bipartite graph. This becomes clear if we factorise the terms of the sum in equation 8:

$$E_{kl} = \sum_{i=1}^{n} \frac{B_{ik}}{\sqrt{k_i}} \frac{B_{il}}{\sqrt{k_i}}. \tag{9}$$

Then we can shortly write $E = D^{\mathrm{T}}D$ with $D_{ik} = B_{ik}/\sqrt{k_i}$ and verify the Euclidean normalisation of the n row vectors of D (for unweighted networks for which we have $B_{ik}^2 = B_{ik}$):

$$\sum_{k=1}^{m} D_{ik}^2 = \sum_{k=1}^{m} \frac{B_{ik}^2}{k_i} = \frac{1}{k_i} \sum_{k=1}^{m} B_{ik} = 1. \tag{10}$$

On the other hand, the projection of the normalised bipartite graph described by affiliation matrix D back on a network of the original nodes is given by DD^{T}. An element of adjacency matrix DD^{T} is given by

$$\sum_{k=1}^{m} D_{ik}D_{jk} = \sum_{k=1}^{m} \frac{B_{ik}B_{jk}}{\sqrt{k_i k_j}} = \frac{A_{ij}}{\sqrt{k_i k_j}}. \tag{11}$$

Thus, Euclidean normalisation of B's row vectors is equivalent to weighting each link in the original (unweighted) network with the geometric mean of its nodes' inverse degrees. The weighted graph described by adjacency matrix E is not the line graph of the unweighted network described by adjacency matrix A but of the network weighted according to equation 11. It depends on the real relations we model with the network whether this is a realistic weighting.

Now we calculate internal connectivity $\tau(L)$ as the sum of internal degrees of vertices in the line graph:

$$\begin{aligned} \tau(L) &= \sum_{k,l=1}^{m} \mu_k(L) E_{kl} \mu_l(L) \\ &= \sum_{k,l=1}^{m} \mu_k(L) \sum_{i=1}^{n} \frac{B_{ik}B_{il}}{k_i} \mu_l(L). \end{aligned} \tag{12}$$

In the same way, we can calculate external connectivity $\sigma(L)$ as the sum of external degrees in the line graph:

$$\begin{aligned} \sigma(L) &= \sum_{k,l=1}^{m} \mu_k(L) E_{kl} (1 - \mu_l(L)). \\ &= \sum_{k,l=1}^{m} \mu_k(L) \sum_{i=1}^{n} \frac{B_{ik}B_{il}}{k_i} (1 - \mu_l(L)). \end{aligned} \tag{13}$$

Now we use the relations

$$\sum_{k=1}^{m} \mu_k(L) B_{ik} = k_i^{\text{in}}(L)$$

and

$$\sum_{l=1}^{m} (1 - \mu_l(L)) B_{il} = k_i^{\text{out}}(L),$$

which directly follow from the definition of the incidence matrix B. Thus, we get

$$\tau(L) = \sum_{i=1}^{n} \frac{(k_i^{\text{in}}(L))^2}{k_i}$$

and

$$\sigma(L) = \sum_{i=1}^{n} \frac{k_i^{\text{in}}(L) k_i^{\text{out}}(L)}{k_i}.$$

From this we easily derive total connectivity of a link-induced subgraph as the sum

$$\tau(L) + \sigma(L) = \sum_{i=1}^{n} k_i^{\text{in}}(L) = k_{\text{in}}(L).$$

B Cost-landscape of the bow-tie graph

For the bow-tie graph we expect two link communities, namely the triangle $\{1, 2, 3\}$ and its complement $\{4, 5, 6\}$, cf. Figure 2 and Evans and Lambiotte (2009). To describe the 2^m different possible subgraphs it is advantageous to make use of the spherical topology of any landscape of subgraphs. Indeed, the cost-function landscape of a graph's subgraphs can be seen as the surface of a globe

- with the whole and the empty graph at the poles,
- with all possible subgraphs of the same size on each circle of latitude, and
- with complementary subgraphs situated at antipodes.

The neighbours of a place in the landscape can be reached by adding an element to the set of nodes (for node-induced subgraphs) or of links (for link-induced subgraphs), respectively, or by deleting an element from this set. That means, there are no direct relations between places on the same circle of latitude. Steps (adding or removing nodes or links) are moves between neighbouring circles of latitude.

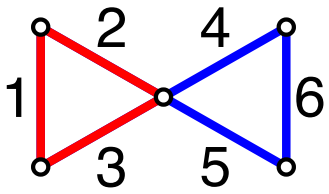

Figure 2: Bow-tie graph with numbered links

We define the north pole as corresponding to the empty subgraph and the south pole as corresponding to the whole graph. The Ψ-globe of the bow-tie graph has five circles of latitude corresponding to six subgraphs with one link, 15 with two, 20 with three, 15 with four, and six with five links, respectively.

For the empty graph at the north pole $\sigma = 0$ and $\Psi = 1$ (by definition). The six single links as the smallest real subgraphs are located at the highest circle of latitude. The two outer links 1 and 6 have $\sigma = 1 \cdot 1/2 + 1 \cdot 1/2 = 1$ and $\Psi = 0.6$, the four inner links have $\sigma = 1 \cdot 1/2 + 1 \cdot 3/4 = 5/4$ and $\Psi = 0.75$.

There are ten connected and five unconnected subgraphs with two links:

- four connected subgraphs with one outer link and one inner link (e.g. link set $\{1, 2\}$) resulting in $\sigma = 1 \cdot 1/2 + 1 \cdot 3/4 = 5/4$ and $\Psi \approx 0.469$,
- six connected subgraphs with two inner links (e.g. link set $\{2, 3\}$) and $\sigma = 1 \cdot 1/2 + 2 \cdot 2/4 + 1 \cdot 1/2 = 2$ and $\Psi = 0.75$,
- four unconnected subgraphs with one outer and one inner link (e.g. link set $\{1, 4\}$) and $\sigma = 9/4$ and $\Psi \approx 0.844$,
- one unconnected subgraph with two outer links ($\{1, 6\}$) and $\sigma = 2$ and $\Psi = 0.75$.

On the equator of the Ψ-globe there are 20 triples of links which can be classified into four types:

- the triangle $\{1, 2, 3\}$ and its complement $\{4, 5, 6\}$ with $\sigma = 2 \cdot 2/4 = 1$ and $\Psi = 1/3$,
- four triples of inner links (e.g. link set $\{2, 3, 4\}$) and their unconnected complements (e.g. link set $\{1, 5, 6\}$) with $\sigma = 3/2 + 3/4 = 9/4$ and $\Psi = 0.75$,

- eight subgraphs with one of the two outer links, one of the two attached inner links and one of the two inner links not attached to the outer link (e.g. link set $\{1, 2, 4\}$): they all have $\sigma = 1 \cdot 1/2 + 2 \cdot 2/4 + 1 \cdot 1/2 = 2$ and $\Psi = 2/3$,
- the unconnected triple with one outer and two inner links (set $\{1, 4, 5\}$) and its unconnected complement (set $\{2, 3, 6\}$) with $\sigma = 3$ and $\Psi = 1$.

On the two circles of latitude on the southern hemisphere we find the complements of the subgraphs on the northern hemisphere with the same Ψ-values. Next to the equator we find 13 connected and two unconnected subgraphs with four links each:

- two unconnected quadruples with one triangle and the second outer link (e.g. link set $\{1, 2, 3, 6\}$) which have $\sigma = 2$ and $\Psi = 0.75$,
- the central star with all four inner links $\{2, 3, 4, 5\}$ with $\sigma = 4/2 = 2$ and $\Psi = 0.75$,
- four subgraphs containing one of the two triangles plus one of the two inner links (e.g. link set $\{1, 2, 3, 5\}$) which all have $\sigma = 1 \cdot 3/4 + 1 \cdot 1/2 = 5/4$ and $\Psi \approx 0.469$,
- the four subgraphs with both outer links and two inner links connecting them (e.g. link set $\{1, 2, 5, 6\}$) with $\sigma = 2/2 + 4/4 = 2$ and $\Psi = 0.75$,
- the four subgraphs with one outer link and three inner links (one of them attached to the outer link, e.g. link set $\{1, 2, 4, 5\}$) with $\sigma = 3/2 + 3/4 = 9/4$ and $\Psi \approx 0.844$.

All complements of the six single links containing the five other links are connected and have the same Ψ-values as their single-link complements (cf. above). The full graph at the south pole with $\sigma = 0$ and $\Psi = 1$ is connected. The Ψ-landscape of links has two local minima: the two triangles have a locally and globally minimal $\Psi = 1/3$. There are no other local minima. Thus, we obtain the pair of complementary triangles as the only solution.

References

Ahn, Y., J. Bagrow, and S. Lehmann (2010). Link communities reveal multiscale complexity in networks. *Nature 466*(7307), 761–764.

Amelio, A. and C. Pizzuti (2014). Overlapping Community Discovery Methods: A Survey. *Social Networks: Analysis and Case Studies*, 105. `http://arxiv.org/abs/1411.3935`.

Clauset, A. (2005). Finding local community structure in networks. *Physical Review E 72*(2), 26132.

Evans, T. and R. Lambiotte (2009). Line graphs, link partitions, and overlapping communities. *Physical Review E 80*(1), 16105.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*(3-5), 75–174.

Gach, O. and J.-K. Hao (2012, January). A memetic algorithm for community detection in complex networks. In C. A. C. Coello, V. Cutello, K. Deb, S. Forrest, G. Nicosia, and M. Pavone (Eds.), *Parallel Problem Solving from Nature - PPSN XII*, Number 7492 in Lecture Notes in Computer Science, pp. 327–336. Springer Berlin Heidelberg.

Gong, M., B. Fu, L. Jiao, and H. Du (2011, November). Memetic algorithm for community detection in networks. *Physical Review E 84*(5), 056101.

Havemann, F., J. Gläser, M. Heinz, and A. Struck (2012). Evaluating overlapping communities with the conductance of their boundary nodes. *arXiv preprint arXiv:1206.3992*.

Havemann, F., M. Heinz, A. Struck, and J. Gläser (2011). Identification of Overlapping Communities and their Hierarchy by Locally Calculating Community-Changing Resolution Levels. *Journal of Statistical Mechanics: Theory and Experiment 2011*, P01023. doi: 10.1088/1742-5468/2011/01/P01023, Arxiv preprint arXiv:1008.1004.

Hric, D., R. K. Darst, and S. Fortunato (2014, December). Community detection in networks: Structural communities versus ground truth. *Physical Review E 90*(6), 062805.

Lancichinetti, A., S. Fortunato, and J. Kertesz (2009). Detecting the overlap-

ping and hierarchical community structure in complex networks. *New Journal of Physics 11*, 033015. arXiv:physics.soc-ph/0802.1218.

Li, Z., X.-S. Zhang, R.-S. Wang, H. Liu, and S. Zhang (2013, December). Discovering link communities in complex networks by an integer programming model and a genetic algorithm. *PLoS ONE 8*(12), e83739.

Ma, L., M. Gong, J. Liu, Q. Cai, and L. Jiao (2014, June). Multi-level learning based memetic algorithm for community detection. *Applied Soft Computing 19*, 121–133.

Merz, P. (2012, January). Memetic algorithms and fitness landscapes in combinatorial optimization. In F. Neri, C. Cotta, and P. Moscato (Eds.), *Handbook of Memetic Algorithms*, Number 379 in Studies in Computational Intelligence, pp. 95–119. Springer Berlin Heidelberg.

Neri, F., C. Cotta, and P. Moscato (Eds.) (2012). *Handbook of Memetic Algorithms*, Volume 379 of *Studies in Computational Intelligence*. Springer, Berlin.

Pan, L., C. Wang, and J. Xie (2012). Link communities detection via local approach. In T. Li, H. Nguyen, G. Wang, J. Grzymala-Busse, R. Janicki, A. Hassanien, and H. Yu (Eds.), *Rough Sets and Knowledge Technology*, Volume 7414 of *Lecture Notes in Computer Science*, pp. 282–291. Springer Berlin / Heidelberg.

Pan, L., C. Wang, J. Xie, and M. Liu (2011). Detecting link communities based on local approach. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, Volume 0, Los Alamitos, CA, USA, pp. 884–886. IEEE Computer Society.

Pizzuti, C. (2009). Overlapped community detection in complex networks. In *GECCO*, Volume 9, pp. 859–866.

Pizzuti, C. (2012). Boosting the detection of modular community structure with genetic algorithms and local search. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 226–231.

Shi, C., Y. Cai, D. Fu, Y. Dong, and B. Wu (2013, September). A link clustering based overlapping community detection algorithm. *Data & Knowledge Engineering 87*, 394–404.

Vitela, J. E. and O. Castaños (2012, May). A sequential niching memetic algorithm for continuous multimodal function optimization. *Applied Mathematics and Computation 218*(17), 8242–8259.

Wei, Y.-C. and C.-K. Cheng (1989, November). Towards efficient hierarchical designs by ratio cut partitioning. In *IEEE International Conference on Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers*, pp. 298–301.

Xie, J., S. Kelley, and B. K. Szymanski (2013, August). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Comput. Surv. 45*(4), 43:1–43:35.

Yang, J. and J. Leskovec (2012). Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, pp. 3. ACM.